CODESSA PRO

COmprehensive **DE**scriptors for **S**tructural
and **S**tatistical **A**nalysis

User's Manual

by

Alan R. Katritzky, Ruslan Petrukhin, Inna Petrukhina, Andre Lomaka and
Douglas B. Tatham
(University of Florida)

Mati Karelson
(Universtity of Tartu, Estonia)

Gainesville, 2001

Contents

INTRODUCTION **2**

CONCEPTS AND DEFINITIONS **3**

STARTING A NEW PROJECT **6**

CREATING STRUCTURE FILES 6
CREATING A PROPERTY FILE 7

WORKSPACE **9**

WORK AREA **11**

STRUCTURE VIEW WINDOW 11
CORRELATION VIEW WINDOW 12
PROPERTY WINDOW 13
LOG WINDOW 14

MANIPULATING LISTS **15**

CALCULATING CORRELATIONS **17**

VIEWING CORRELATIONS **17**

Introduction

CODESSA PRO is an entirely new software package, which performs tasks similar to CODESSA, but with many distinct advantages over the previous software package. In particular, its add-in mechanism makes CODESSA PRO expandable, its calculation engine has been optimized at the assembly language level for Pentium, Pentium Pro, Pentium II, Pentium III and Pentium IV processors, and finally, it was designed to run in 32-bit Windows environments. CODESSA (Comprehensive Descriptors for Structural and Statistical Analysis) is a comprehensive program for developing quantitative structure/property relationships (QSPR), which integrates all necessary mathematical and computational tools to calculate a large variety of molecular descriptors on the basis of the 3D geometrical and/or quantum-chemical structural input of chemical compounds, to develop (multi)linear and non-linear QSPR models of the chemical, physical or biological properties of individual compounds, to perform cluster analyses of the experimental data and molecular descriptors, to interpret the developed models, and to predict properties for compounds previously unknown or unavailable.

This manual presents the guidelines for a successful development of QSAR/QSPR models using the CODESSA PRO program. The execution of each step of the program normally requires the knowledge about the previous proceedings, and thus they are discussed in the sequence you would accomplish them. This manual also provides further information concerning the features and methods available in the CODESSA PRO program, their purpose, and interpretation of the results. For a more detailed description of the program and employed techniques please refer to the CODESSA PRO *User's Manual* and *Reference Manual*.

The CODESSA program is designed to operate in the following Microsoft Windows environments: Windows 2000, Windows NT, and Windows 9x. To start CODESSA double click on its icon shown above. When the program starts, the CODESSA PRO Visual Interface (CVI) window will open. Click **View** on the menu-bar to open a pull-down menu, next click on the **Refresh** option to refresh the current snapshot, which is then displayed on the screen. Alternatively, the single keystroke **F5** will refresh the snapshot. Before attempting to use CODESSA PRO it is necessary to understand some of the concepts (terminology) used by the program.

Concepts and Definitions

The use of CODESSA PRO requires understanding of the following concepts:

Artifacts:

The individual files or items used by CODESSA PRO. There are four types of artifacts: structures, descriptors, properties, and correlations. Using the CODESSA PRO program you will frequently manipulate single artifacts or lists of artifacts. The names of properties are declared in the property (prp) file and each must have a uniquely defined name. CODESSA PRO names most artifacts (*e.g.* descriptors and correlations). Structures are currently named by the operator, but when the CMol3D module is complete, it will automatically create and name structures.

Structure:

A representation of an individual chemical object with a precise chemical constitution. Examples of structures include a single molecule, a monomeric unit of a polymer, or a molecular complex of a definite composition. The minimum information that a structure must include is the types of atoms involved and their connectivity. Each structure must be linked with three files containing: 3-D structure, SCF, and Force information. Before a 3-dimensional structure can be input in CODESSA PRO it must be converted to MDL *molfile* format and be optimized with a gradient norm of 0.01 or better. The *molfile* structures are stored in the mol3dmop directory, SCF output files are stored in the mopscf directory, and Force output files are stored in the moptherm directory. SCF and Force structures will be created and properly stored automatically by the CMol3D module. See **Starting a New Project** for instructions to create each file type. It is vital that the name for a particular structure be exactly the same in each directory. Structure names must be in the form S0000001.xxx and should be added in sequential order (e.g. S0000001.xxx, S0000002.xxx…).

Storage:

All items that are connected with a project are stored in a single location within the file system - in the storage. To change a storage location, click **Option** on the main menu-bar and select **Storage**. The box that opens allows the user to control the storage location of structure files, correlations, descriptors, lists…

Snapshot (Cache File):

The snapshot (CODESSA PRO cache file) is a binary (compressed) representation of all items in storage. This file located in the memory whenever work is being performed on the CVI module. The cache file reloads whenever the CVI detects a change in the storage contents. The snapshot can be reloaded manually anytime by opening the pull-down menu **View** on the menu-bar, then selecting **Refresh**, or using **F5** keyboard shortcut.

Workspace:

The workspace is a window within the CVI that represents the snapshot and located in the workspace area on the top left side of the CVI frame.

Folder:

A folder is used to store lists and is similar to a directory. CODESSA PRO uses four folders: structure, descriptor, property, and model (correlation). Each of the first three folders has a system-generated list named **All**. This list will contain "ALL" of the structures, descriptors, properties, or models for the corresponding folder. The Descriptors folder also includes system defined descriptor lists according to group and type. A particular descriptor Item may be contained in several lists. The Correlations folder contains lists for each property item in the property folder. The list is generated automatically by the system when a property item is created. When correlations are calculated, 50 correlation items (by default) will be stored in the list that corresponds to the property.

List:

A list is collection of the chemical structures, descriptors, physico-chemical properties, or models (correlations). Lists can be either system type or user type. System lists cannot be deleted by the user.

Current Analysis Dimensions:

The dimensions used for an analysis by CODESSA PRO are the property, the list of descriptors, and the list of the structures. All the dimensions are selected from **Dimensions** pull-down menu. The default dimensions for a new snapshot file are:

Dimension	Default value
Property	Not selected
List of descriptors	**All** group
List of structures	**All** group

The selected dimensions are used in the formation of the descriptor/property matrix.

Descriptor/Property Matrix:

The descriptor/property matrix consists of descriptor and property (the last column) values. The horizontal dimension of the matrix is descriptor/property ID sequence and the vertical dimension is the structure ID sequence. The matrix has two presentations: binary and text. The binary presentation is used for internal use, while the text presentation is optimized for import into the STATISTICA software package.

Property:

A property is a physical or chemical characteristic, biological activity or other characteristic of concern. Each property value must be associated with a **structure** located in the Structure folder.

Descriptors:

Descriptors are defined as numerical characteristics associated with chemical structures. They are derived on the basis of the structures chemical constitution, topology, geometry, and inherent wavefunction and potential energy surface. The values of a particular **descriptor** can be provided by the user or calculated by the CODESSA program. Each **descriptor** value must be associated with a previously defined **structure**. Descriptors calculated by the CODESSA program are named automatically; renaming descriptors is not recommended.

Correlation:

A correlation represents the results of the (multi)linear regression between a property of interest (y) and one or more selected descriptors (x_i).

$$y = a_0 + \sum a_i x_i$$

Correlations are composed of regression coefficients (a_i), correlation coefficient (R^2), standard error s^2, and Fisher criterion F value for the set of structures used in its derivation. By default each **correlation** is named by its number of structures (N), correlation coefficient (R^2), crossvalidated correlation (R^2_{CV}) and Fisher criterion value (F), and standard error (s^2).

Artifacts:

Items used by or created by CODESSA PRO. The four types of artifacts are structure, property, descriptor, and correlation.

Starting a new project

It is important to decide whether to use a general storage area for many projects or separate storage areas for each individual project. A general storage area allows a research group to avoid repeating the calculations of structures and descriptors as well as minimizing the amount of storage space used. If a general storage area is to be used, it will be necessary to keep an index of the structure and molecule names. To prevent confusion exact IUPAC names should be used. Additionally, the names of the molecules should be inserted in the name section of the *molfile* by opening the file with a text editor and entering the name on the first line and comments on the third line.

Creating structure files

Molfiles

Structures can be created using a chemical drawing program that will convert 2-D structures to 3-D structures, i.e. HyperChem, ISIS Draw, ChemDraw, or PCModel. The 3-D geometries must be optimized in a two-step process using MOPAC. Each step involves adding appropriate keywords to the structure file, running MOPAC, and then converting the optimized output file into a MOPAC internal file. The conversion can be performed manually or by using a utility such as Babel (http://www.ccl.net/cca/software/MS-WIN95-NT/babel/index.shtml). The first optimization uses the keyword: AM1 and the second optimization uses the keywords: AM1 GNORM=0.01 PRECISE. The additional keyword MMOK should be used for any molecule with peptide linkages. If either optimization fails, it should be repeated using additional keywords until successful. The following list contains the progression of additional keywords to add.
EF
XYZ
EF XYZ
EF XYZ GEO-OK
If none of the keywords or combinations above work, select a different starting point for the z-matrix, rearrange atoms accordingly, or add a dummy atom and create a new z-matrix. Dummy atoms must be deleted prior to performing SCF and thermodynamic calculations.
The optimized structures should be converted to molfile format, given the extension mol, and stored in the mol3dmop directory. The first line of the molfile will be blank after MOPAC optimization, insert the name of the molecule in the first line and comments on the third line.

SCF files

Convert the *molfile* structures to MOPAC internal files and add the keywords: AM1 VECTORS BONDS PI POLAR PRECISE ENPART EF. Perform the MOPAC calculation and save the files with an mno extension in the mopscf directory.

Force Files

Convert the *molfile* structures to MOPAC internal files and add the keywords: AM1 FORCE PRECISE THERMO ROT=1. Perform the MOPAC calculation and save the files with an mno extension in the moptherm directory.

Creating a property file

Once the property value of interest has been collected and carefully checked for each structure the data must be assembled in a text (prp) file. Spreadsheet programs may be used to simplify the process of assembling data, but the file should be saved as text. The format of the file must be exactly as follows:

Name of property file
Comments
Structure1 Value1
Structure2 Value2
Structure3 Value3

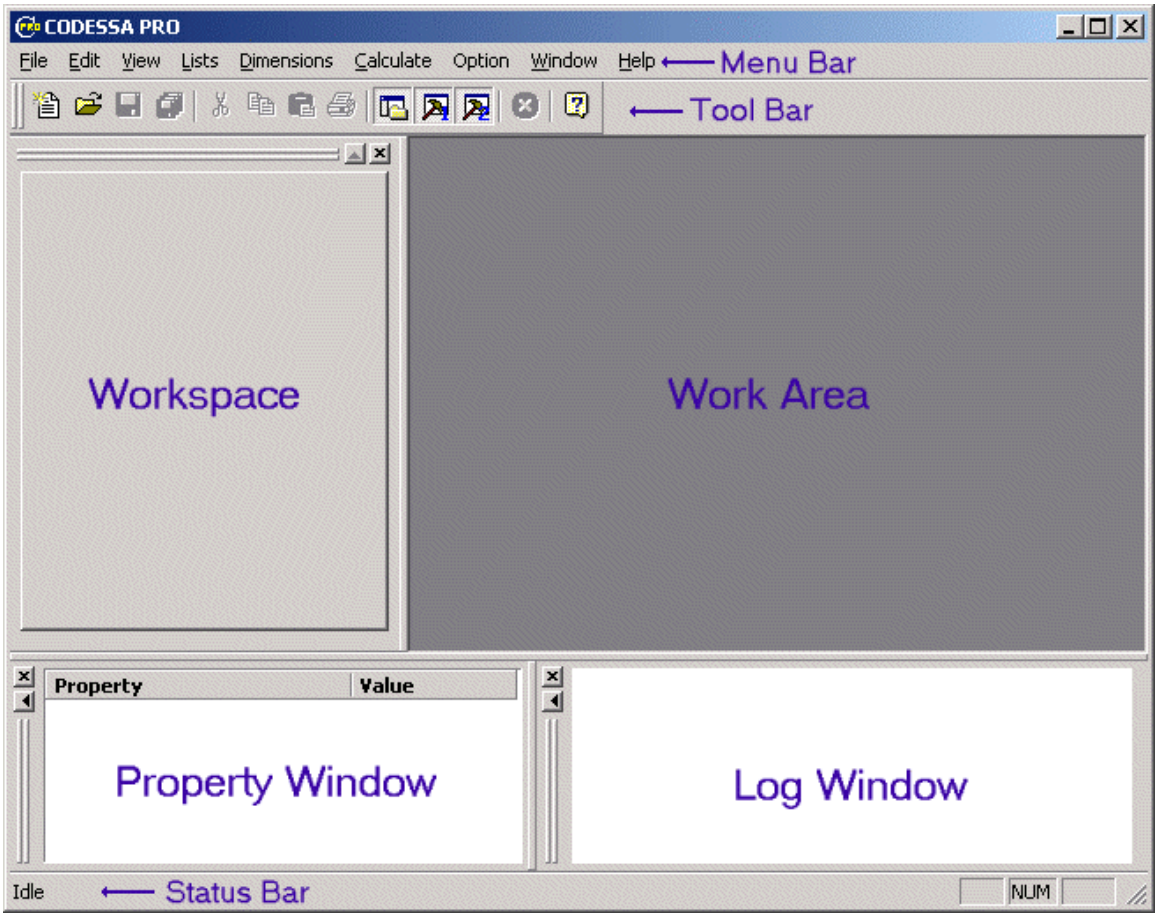
Name of property file: The name that will be assigned to the property artifact. The artifact will be contained within one or more lists in the property folder. A list with same name will also be generated automatically by CODESSA PRO and stored in the correlations folder.

Comments: Any comments written in this location will be listed in the comments section of the property window whenever the property is selected.

Structure1: The number of the first structure (not the filename). Proceeding 0's in the structures number are ignored; the structure S0000023.mol could be listed as 00023, 023, or just 23.

Value1: The property value associated with structure1. There must be white space (space or tab) between the structure number and the value. The type or number of white spaces is irrelevant so long as some type of space exists.

CODESSA PRO Visual Interface



Workspace: An on-screen presentation of the current cache file.

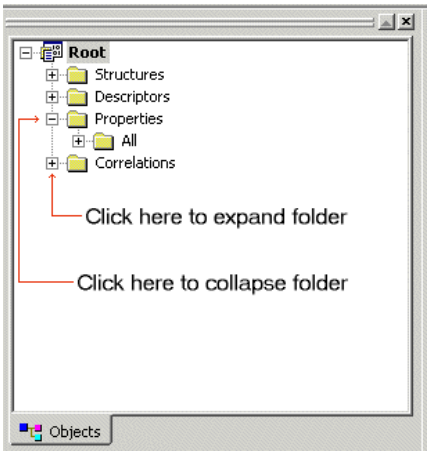
Work Area: The work area the screen space for various view windows.

Property Window: The property window depicts information about properties of the selected object and available for almost all objects.

Property Window: The window purpose is protocol all operation on the storage.

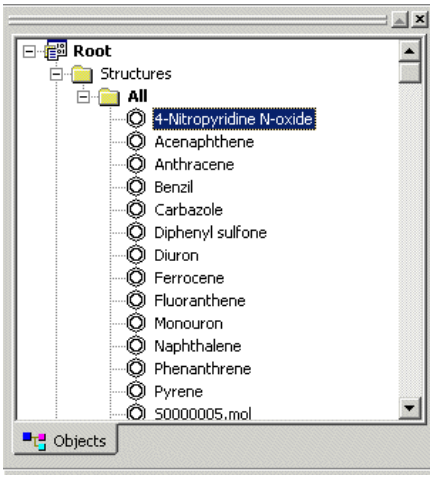
Tool Bar: Shortcuts to commonly performed tasks. The task will be displayed it the pointer is positioned over the icon.

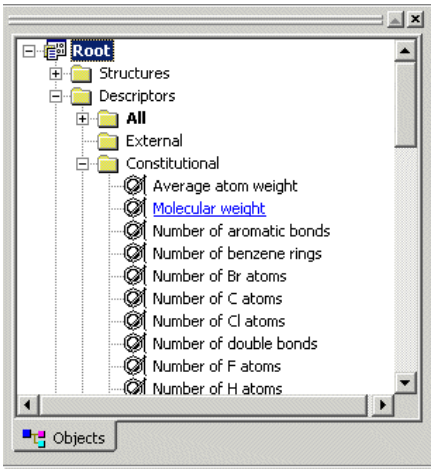
Workspace



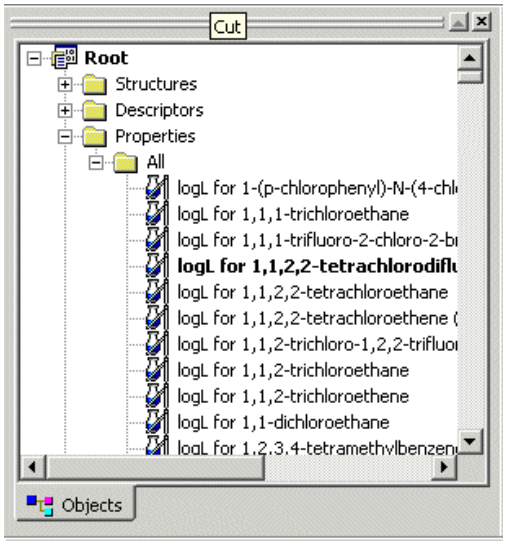
The workspace area contains an on-screen presentation of the current cache file which is essentially a directory tree. Clicking on the minus will collapse a folder (branch of the tree) and clicking on a plus will expand a folder. When a folder is expanded (note the **Properties** folder) it will show the lists (note **All** list) contained in the folder. The same icon is used to represent lists as folders because they are also used to contain files. If a list is empty then there will be no plus or minus beside it.

The **Structures** folder is expanded to show the list **All**. The list **All** is also expanded to show the structure artifacts. Notice the icon beside the structure artifact. Each type of artifact will have a different icon. The name that is listed will be the same as the name given on the first line of the molfile. The last structure artifact gives the filename of the molfile because the name was not entered on the first line of the molfile. Double clicking on an artifact launches a **structure view** window in the **work area**.



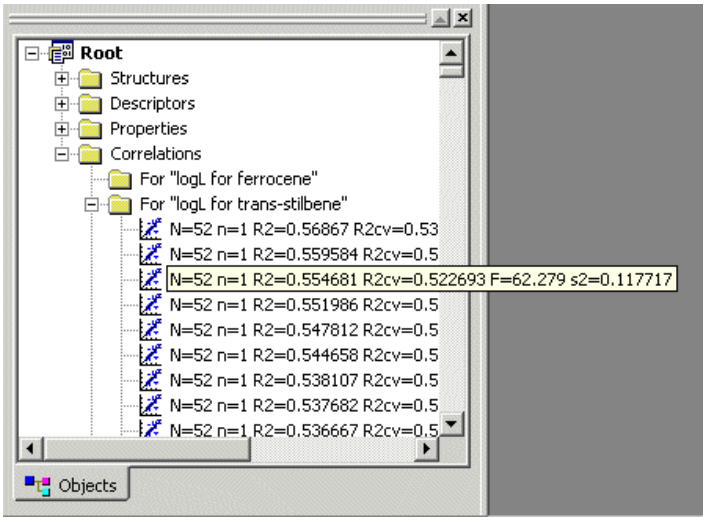


The **descriptors** folder is expanded to show the **All**, **External**, and **Constitutional** lists as well as several descriptor artifacts. The **All** list is in boldface font because it is the currently selected analysis dimension for descriptors. Notice the descriptor icon is different from the structure icons pictured above. If an artifact is selected its color will change from black to blue as can be seen with the Molecular Weight artifact.



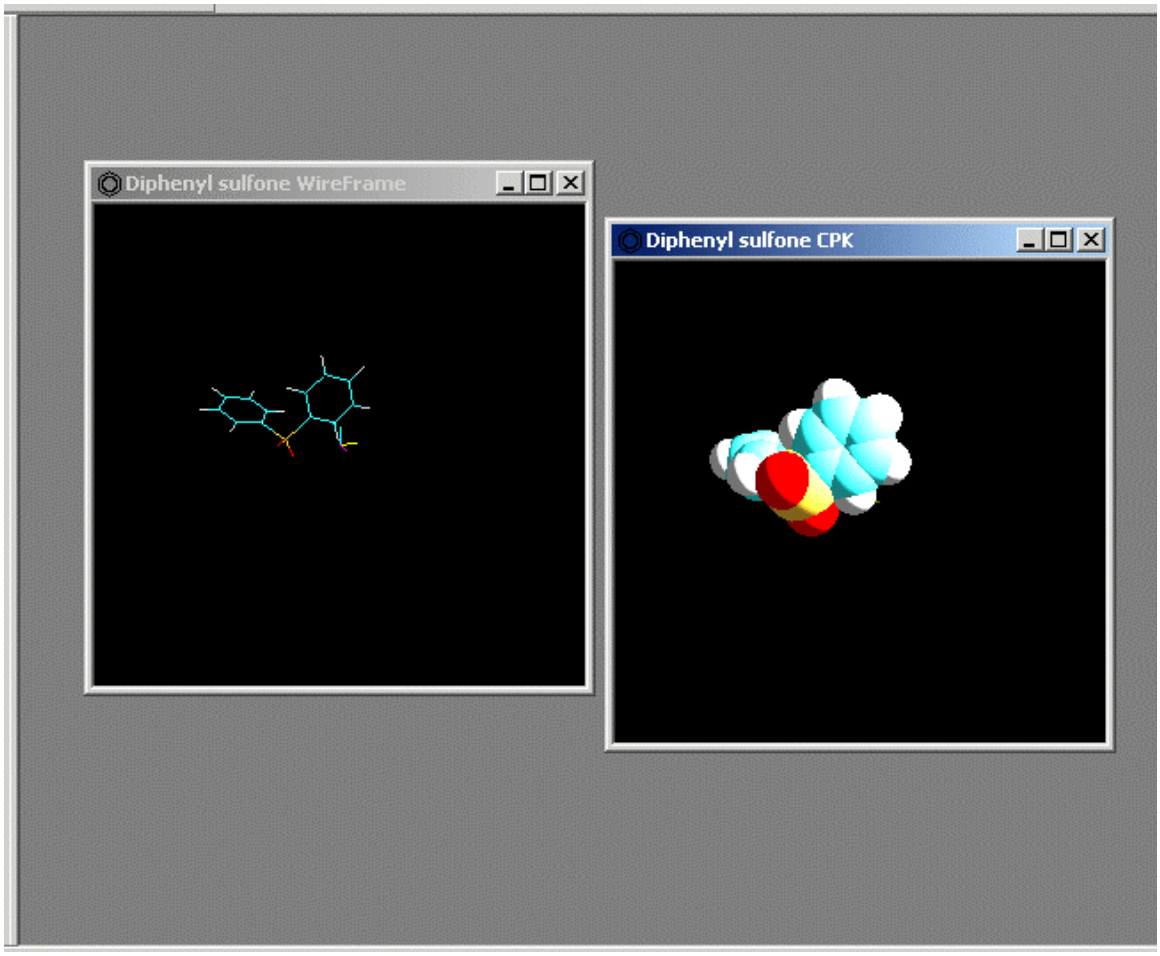
The **Properties** folder is expanded to show the **All** list and several property artifacts. Notice that the forth property is in boldface font, which indicates that it is the currently selected property dimension. Notice the icon used to represent property artifacts. The name of the property artifact is the same as the first line of the property (P0000004.prp) file. The artifacts are created automatically for each property file stored in the props subdirectory.

The correlations folder has been expanded to display several lists and artifacts. A list is created automatically for each property artifact. Notice there is no plus or minus beside the first list. This indicates that no model for the property has been calculated and the list contains on artifacts. If the pointer is held over a correlation artifact, the details of the artifact are displayed. Double clicking on an artifact launches **correlation view**.



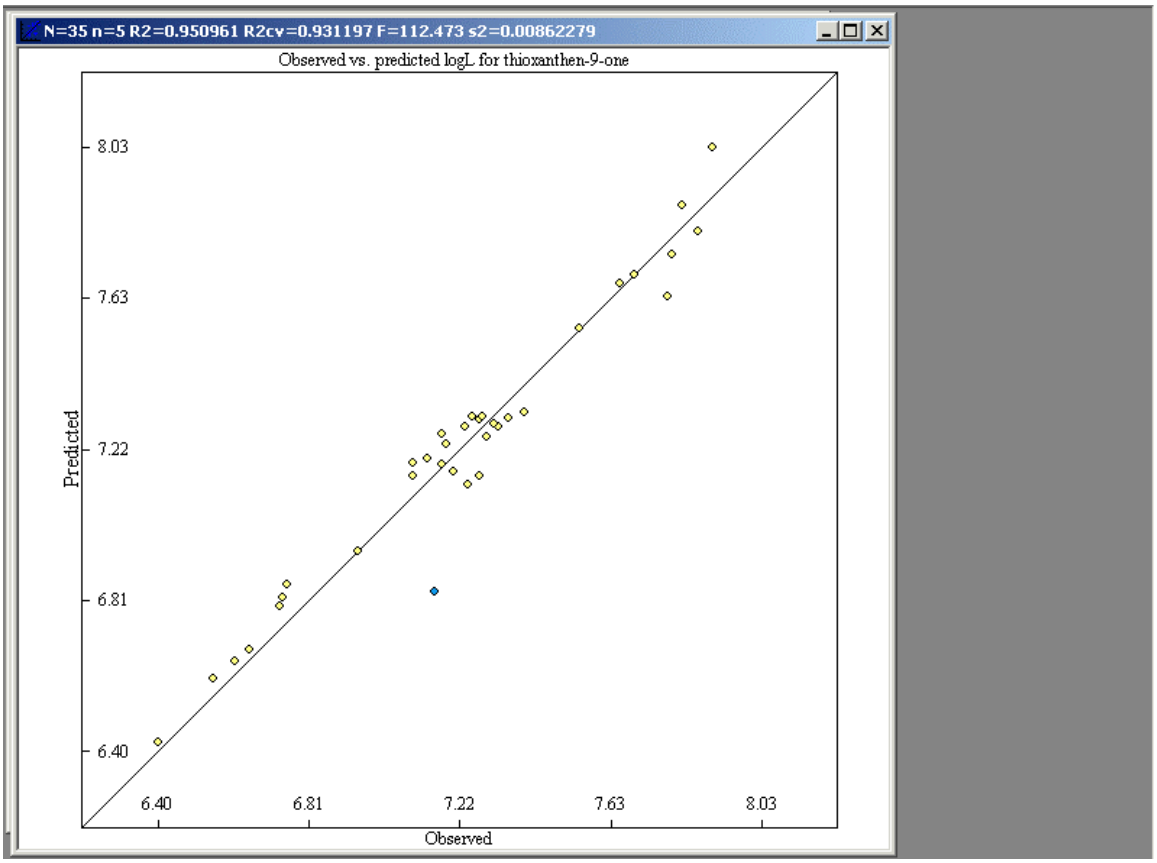
Work Area

Structure View Window



Double clicking on a structure artifact in the structures folder opens a **structure view window** in the work area. Structures are 3-dimensional and can be viewed as wire-frame, ball and stick, CPK surface filled, or solvent accessible surface. Right clicking inside the window opens the view and label context pop-up menu. Selecting view provides the four options mentioned earlier and label identifies the number or each atom or the type atom. When the pointer is positioned over the view window it changes to a 4-sided arrow that can be dragged to rotate the structure.

Correlation View Window

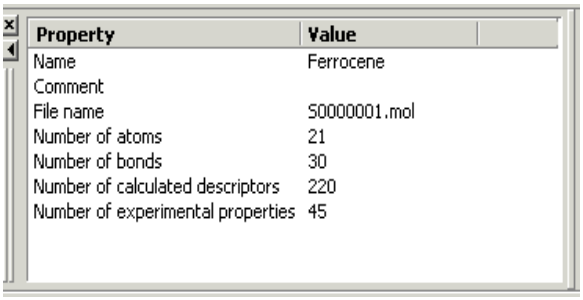


Double clicking on a correlation artifact opens the **correlation view window**. The graph shows the observed (experimental) values versus the predicted values for the property. When the window is first opened the points for all outliers will be blue rather than yellow. The color of the points will change to blue to indicate they have been selected. Clicking once on a point will cause its properties and their associated values to be displayed in the properties window. Double clicking on a point will cause a **structure view window** for the structure that corresponds the point to open in the work area. Right clicking side the window will open a context pop-up menu that will open the structure view windows for each of the structures used in the correlation, the correlations for each descriptor, allow the creation of lists, and allow marking of structures.

Double clicking on a descriptor artifact will also open a **correlation view window**. The window will show the relationship between the descriptor and the selected property dimension.

Property Window

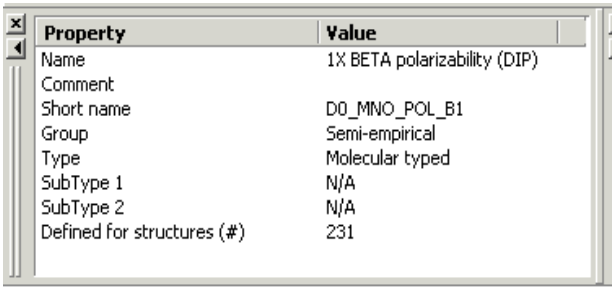
The item selected determines the contents of the property window. If a folder or list is selected in the workspace, the number of artifacts contained is given. If an artifact is selected it provides varying information, depending on the type of artifact.



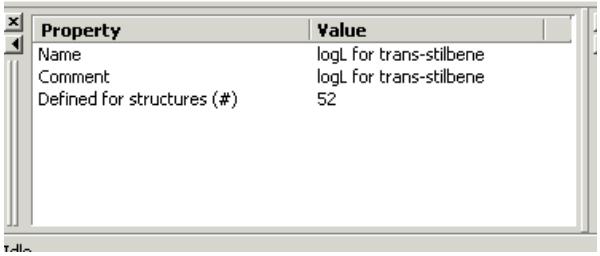
Property	Value
Name	Ferrocene
Comment	
File name	S0000001.mol
Number of atoms	21
Number of bonds	30
Number of calculated descriptors	220
Number of experimental properties	45

The picture on the left shows information given in the property window when a structure artifact is selected.

Pictured is the information provided when a descriptor artifact is selected in the workspace.



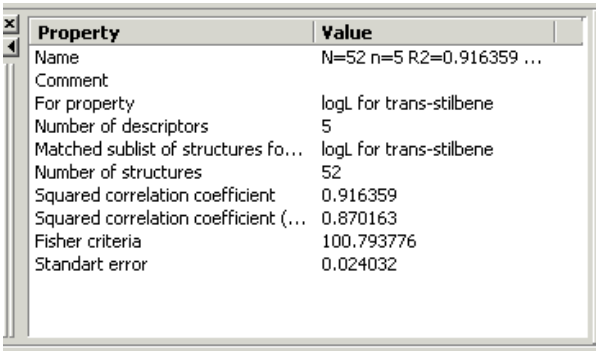
Property	Value
Name	1X BETA polarizability (DIP)
Comment	
Short name	D0_MNO_POL_B1
Group	Semi-empirical
Type	Molecular typed
SubType 1	N/A
SubType 2	N/A
Defined for structures (#)	231



Property	Value
Name	logL for trans-stilbene
Comment	logL for trans-stilbene
Defined for structures (#)	52

When a property artifact is selected the information from the prp file is shown, this includes name, comments and the number of structures.

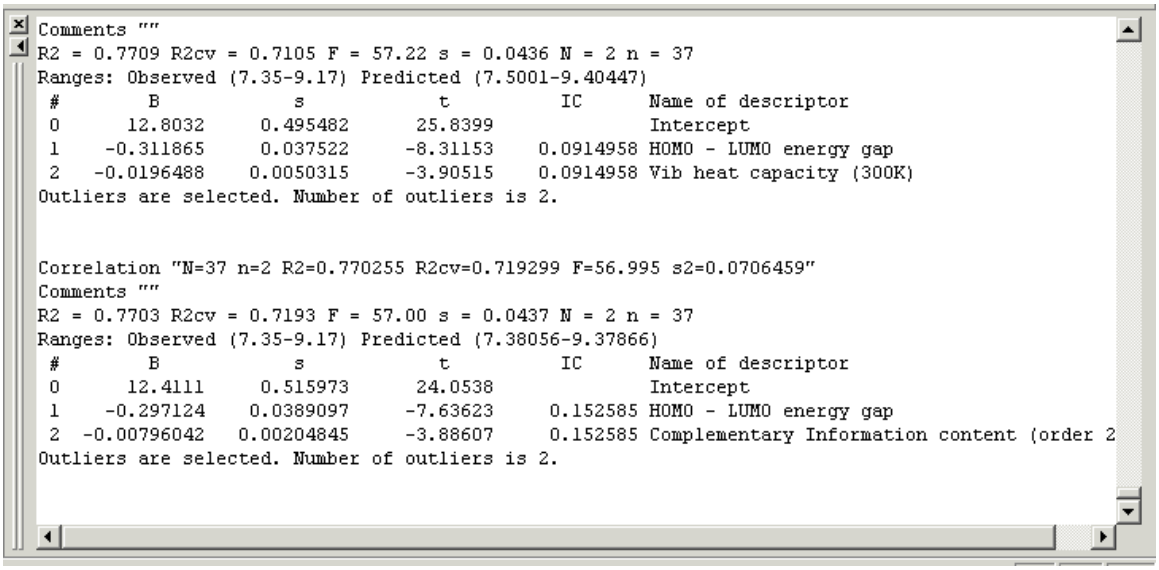
The property window for correlation artifacts lists the details of the correlation. Details include the properties used in the name of the correlation and the property name.



Property	Value
Name	N=52 n=5 R2=0.916359 ...
Comment	
For property	logL for trans-stilbene
Number of descriptors	5
Matched sublist of structures fo...	logL for trans-stilbene
Number of structures	52
Squared correlation coefficient	0.916359
Squared correlation coefficient (...	0.870163
Fisher criteria	100.793776
Standart error	0.024032

Log Window

The log window describes each operation that has been performed on the storage. Double clicking on a correlation in the workspace will give a full description of the correlation (as pictured below). The text is in RTF format and can be cut and pasted directly into a word processing program.



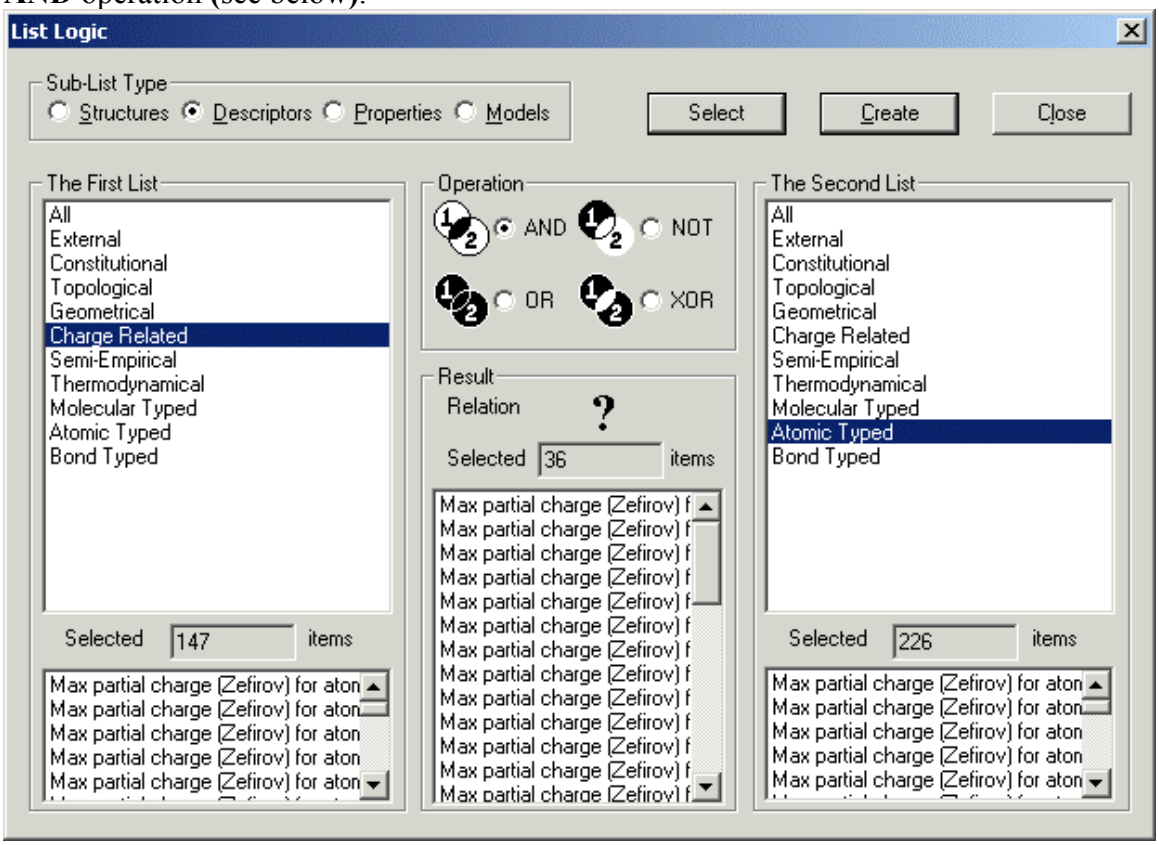
```
Comments ""
R2 = 0.7709 R2cv = 0.7105 F = 57.22 s = 0.0436 N = 2 n = 37
Ranges: Observed (7.35-9.17) Predicted (7.5001-9.40447)
 #        B            s            t          IC      Name of descriptor
 0      12.8032     0.495482      25.8399              Intercept
 1    -0.311865     0.037522     -8.31153    0.0914958 HOMO - LUMO energy gap
 2   -0.0196488    0.0050315     -3.90515    0.0914958 Vib heat capacity (300K)
Outliers are selected. Number of outliers is 2.


Correlation "N=37 n=2 R2=0.770255 R2cv=0.719299 F=56.995 s2=0.0706459"
Comments ""
R2 = 0.7703 R2cv = 0.7193 F = 57.00 s = 0.0437 N = 2 n = 37
Ranges: Observed (7.35-9.17) Predicted (7.38056-9.37866)
 #        B            s            t          IC      Name of descriptor
 0      12.4111     0.515973      24.0538              Intercept
 1    -0.297124    0.0389097     -7.63623     0.152585 HOMO - LUMO energy gap
 2  -0.00796042   0.00204845     -3.88607     0.152585 Complementary Information content (order 2
Outliers are selected. Number of outliers is 2.
```

The edit pull-down menu has options to **clear all log**, **paste to log**, **select all log**, and **copy from log**. Transferring the information for several correlations to a word processing program can be accomplished by clearing the log, double clicking on each of the correlation, clicking select all, then copy.

Manipulating Lists

CODESSA PRO automatically one or more lists in each folder, but when analyzing a property or correlation it is sometimes helpful to create a user-defined list, e.g. a list of structures containing phenyl rings. The simplest method for creating a list in CODESSA PRO is to select several artifacts from one or more existing lists (i.e. **all**), right click to open the context pop-up menu and left click on create list. A list can also be formed from a group of artifacts that are linked to a *common* artifact. If a descriptor is chosen to be the *common* artifact, then all the structures that are defined for that descriptor could be selected for a list. To create the list from a *common* artifact, select the *common* artifact, open the context pop-up menu, click on **Select Linked**, and then select the type of artifacts that you want to add to the list. If multiple *common* artifacts are selected the list that is created corresponds to creating lists for each *common* artifact, then performing the **AND** operation **(**see below**)**.



The **List Logic** module in CODESSA PRO compares the artifacts contained in two lists and can be opened by clicking the pull-down menu **List** then left clicking on **Logic**. Choose the type of artifact that will be contained by the list by clicking on the appropriate button in the **Sub-List Type** box (**descriptors** is selected in the example above). Choose the desired lists for comparison and the appropriate **Operation**. The artifacts from the two lists that meet the criteria of the operation will be displayed in the **Results** box. The results can then be copied to a new list (**Create**) or be selected (**Select**) for further comparisons. When comparing three or more lists, first compare two of the lists, click on **Select**, restart **List Logic**, select **Current Selection** in **The First List** box, then select the remaining list in **The Second List** box.

Calculating Descriptors

One of the major advantages of CODESSA is its tremendous pool of *descriptors*, which are calculated for each of the structures listed in the prp file. The *descriptors* are divided into ten lists. The first list contains ***external*** descriptors. The remaining lists can be divided into two categories. The first category divides descriptors into lists based on the type of calculation, while the second category divides the descriptors into lists according to the portion of the molecule being described.
The first category contains the following lists: ***constitutional, topological, geometrical, charge related, semi-empirical***, and ***thermodynamic***. The second contains the remaining lists: ***molecular type, atomic type,*** and ***bond type***. For more detailed information on descriptors available in the program we refer to the CODESSA PRO *Reference Manual*.
At this point it is assumed that the structures and property files have been saved in the appropriate directories. Calculation of descriptors is considerably faster than with the previous versions of CODESSA.

Calculating descriptors in CODESSA is simple

1. Expand the list in the properties folder that contains the property artifact of interest.
2. Select the property artifact (it should be highlighted).
3. Click the **Dimensions** pull-down menu and click on **property** (the property artifact should now be in boldface).
4. The default setting is to calculate **all** descriptors for **all** structures listed in the prp file. Skip to step 7 if using default settings.
5. To calculate subset of the possible set of descriptors, select the (automatically generated or user defined) descriptors list, click the **Dimensions** pull-down menu, then **List of descriptors**.
6. To calculate descriptors for a subset of structures select the user defined structures list, click the **Dimensions** pull-down menu, then **List of structures**.
7. Click the **Calculate** pull-down menu, then **Descriptors**.

The order for the **Calculate** commands is **Descriptors**, **Form matrix**, then **HMPRO**. If **HMPRO** is selected without first running the prior commands, CODESSA PRO will automatically perform the preceding operations.

Calculating Correlations

Calculating correlations in CODESSA PRO is even easier than calculating descriptors.

1. Ensure that property of interest is in boldface. If not, select the property so that it is highlighted and click the **Dimensions** pull-down menu, then **property**.
2. Click on the **Calculate** pull-down menu, and then **Form matrix**.
3. Click on the **Calculate** pull-down menu, and then **HMPRO**.

Viewing Correlations

1. Expand the Correlations folder.
2. Expand the correlation list with the same name as the property of interest. The default number of correlation artifacts in each correlation list is 50, but this can be altered by editing the settings in the HMPRO.par file.
3. Double click on a correlation artifact to open the **Correlation view window** in the work area. A description of the correlation will be displayed in the property window with the complete information in the property log.